基于节点综合相似度的多标签传播社区划分算法 *

郝梓琳，李　雷，施化吉

(江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013)

摘　要：为了解决现有的多标签传播社区划分算法采用的随机顺序策略导致形成的社区划分结果不稳定和社区质量不够高的问题，提出了一种基于节点综合相似度的多标签传播社区划分算法 MLPA-NCS。以节点潜在影响力的降序作为节点选择顺序，解决社区结果划分不稳定问题。根据节点的主题相似度和链接相关度计算出节点综合相似度，并以节点综合相似度降序作为更新节点标签时对邻近节点遍历的顺序，提高所划分社区的质量。采用真实数据集和人工网络数据，对多个算法进行对比实验，结果表明算法有效可行，社区划分结果更稳定，社区质量也更高。

关键词：社区划分；标签传播；重叠社区；综合相似度；主题相似度

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.12.0809

Multi-label propagation algorithm for community division based on node comprehensive similarity

Hao Zilin, Li Lei, Shi Huaji

(*School of Computer Science & Communication Engineering Jiangsu University*, *Zhenjiang Jiangsu* 212013, *China*)

Abstract: To solve the problem that recent research about multi label propagation community division algorithm adopted the random sequence strategy to result in unstable result of community division and poor community quality, this paper proposed a Multi-label Propagation Algorithm Based on the Node Comprehensive Similarity (MLPA-NCS) for community division. This paper chose the descending order of node potential impact as the node selection order in order to solve the problem of the instability of the propagation. Node synthesis similarity could be calculated based on the theme of node similarity and link correlation, and it's descending order was used as the order of neighboring nodes traversal when updating the node label to improve the quality of the communities found. This paper used real data sets and artificial network data to compare the results of several algorithms. The results show that the algorithm is effective and feasible and able to make the result of community division more stable while the quality of community more effectively.

Key words: community division; label propagation; overlapping community; comprehensive similarity; topic similarity

0　引言

社区结构是社交网络中极为重要的特性之一，通常认为同一个社区内关系紧密，社区之间关系稀疏。对社区结构进行划分有助于节省资源，例如对以社区为单位的结构进行广告投放、信息推荐和舆情控制等。

目前已有研究者提出基于标签传播思想对社区进行划分，Reghavan 等人[1]首次提出基于标签传播的社区划分算法 LPA，由于其时间复杂度几乎是线性且算法设计较为简单，使得该算法被广泛应用于大型网络社区划分中，该算法仅允许每个节点拥有一个标签，生成的社区为非重叠社区，然而现实的社交网络中常常有节点同时属于不同的社区，形成较为复杂的重叠社区。为此，很多学者对其进行改进，提出了各种改进算法[2]。其中 Gregory[3]对 LPA 算法进行扩展提出基于多标签传播的重叠社区划分算法 COPRA，允许节点同时携带多个社区的标签和相应标签的隶属度，在重叠社区划分中取得了较为明显的效果，但 COPRA 算法是采用随机顺序策略选择节点更新标签，且在迭代过程中也随机遍历邻接节点的标签集，随机顺序策略的采用使得传播过程不确定，导致社区划分结果不稳定且生成社区质量不够高。Xie 等人[4]则提出了 SLPA 算法来实现重叠社区的划分，节点在每次标签传播中只能够传播一个标签，但是允许节点保留其所有感兴趣的标签，最后统计标签序列中出现的各标签概率。2016 年，刘世超等人[5]提出基于标签传播概率的重叠社区发现算法 LPPB，综合利用网络的结构特点与节点属性

收稿日期：2017-12-18；**修回日期：**2018-01-30　　**基金项目：**江苏省六大人才高峰项目（2014-WLW-012）；江苏省重点研发计划（社会发展）资助项目（BE2016630，BE2015617）

作者简介：郝梓琳（1992-），女，山西晋中人，硕士研究生，主要研究方向为社会网络分析、数据挖掘（zlin_hao@163.com）；李雷（1976-），男，河南南阳人，讲师，主要研究方向为智能信息处理、社会网络分析；施化吉（1964-），浙江台州人，教授，主要研究方向为智能信息处理、社会网络分析.

计算标签传播的概率；张昌理等人[6]提出基于信息熵和局部相关性的多标签传播重叠社区发现算法，按照标签熵从小到大的顺序进行标签更新，使得最后的划分结果相对稳定一些。文献[7]提出了基于边界节点和标签传播的社区划分算法，在一定程度上降低标签传播的随机性。

为此，本文以 COPRA 算法框架为基础对其改进，提出基于节点综合相似度的多标签传播社区划分算法（Multi-label Propagation Algorithm Based on the Node Comprehensive Similarity，MLPA-NCS）。首先以用户节点的潜在影响力降序作为节点更新顺序，以解决社区结果划分不稳定问题。然后考虑节点之间潜藏的主题相似因素和链接关系，以节点主题相似度和链接相关度作为更新节点标签时对邻近节点遍历的顺序，避免因随机策略更新标签带来的不稳定问题，并提高生成社区的质量。

1　相关工作与问题提出

标签传播思想为每个节点赋予唯一的标签，通过在迭代更新中接受邻接节点标签的影响来改变自身的标签，直至标签不再改变，此时标签相同的节点划为同一个社区。COPRA 算法允许同一个节点在迭代更新中携带多个标签，从而使得迭代结束后同一节点可属于多个社区，得到重叠社区结构。

COPRA 算法的基本思想如下：

a）初始化节点标签。初始时对网络中的每一个节点 v 各自赋予不同标签 c，表示其从属的社区，标签对应的隶属度为 b，b 都为 1。以后随着标签传播过程将更新标签节点的标签集，每一个标签有对应的隶属度。

b）标签传播过程。网络中的任意节点 v 在标签传播过程中通过接受邻接节点 u 的影响来更新自己的标签集 $lable(v)$，如此进行迭代更新。在 t 轮迭代中按式（1）计算节点 v 的每一个标签 c 的隶属度。

$$b_t(c,v)=\frac{\sum_{u\in N(v)} b_{t-1}(c,u)}{|N(v)|} \tag{1}$$

并对所有标签的隶属度进行标准化，使隶属度之和为 1，如式（2）所示。

$$\sum_{c\in label(v)} b(c,v)=1 \tag{2}$$

c）直至每个节点的标签集不再更改或者满足迭代次数后停止迭代，标签传播过程停止。

d）根据节点最终的标签集确定其所属的社区。

COPRA 算法按照随机顺序选择网络中节点迭代更新其标签，所采用的同步更新策略会导致下一轮迭代对节点标签的更新依赖于上一轮更新的结果，选择不同的节点更新标签势必会产生不同的划分结果。多次实验发现，社区划分在迭代更新节点标签时对节点的选择顺序非常敏感，通过随机顺序选择节点更新标签会造成每次实验的收敛结果都不一致，且社区划分结果都存在着一定程度的差异。实际上网络中影响力大的节点通常也是网络中的重要节点，PageRank 中心性认为节点的重要性取决于邻接节点的度及其重要性，若节点的邻接节点在网络中很重要，则该节点成为重要节点的可能性也越大。因此可以用 PageRank 中心性评估节点的潜在影响力，按其降序选择节点进行标签更新，可以在一定程度上解决由于随机顺序选择节点造成的划分结果不稳定问题并提高划分社区的质量。

COPRA 算法在更新节点标签时只是随机遍历其邻接节点的标签对其影响以更新自己的标签集，且忽略了不同邻近节点（分为直接邻接节点和间接邻接节点，其含义见后文 2.1.2 节解释）对其影响程度的差异。在实际社交网络中，尽管很多用户之间没有直接邻接，但是他们拥有相同的粉丝，或者共同关注某人，这说明他们之间在一定程度上也有间接的联系或影响，因此在更新节点标签时，不仅要考虑直接邻接节点而且还要考虑间接邻接节点的影响，为此本文通过节点链接相关度来度量节点间的链接关系。除了链接关系外，社交网络中的用户之间在关注的主题上也存在一定的相似性，通过用户的主题相似性程度进行社区划分，所得到的社区更具有相对一致的主题，得到的社区质量更高。

2　基于节点综合相似度的多标签传播算法 MLPA-NCS

MLPA-NCS 算法首先初始化节点标签，然后依据节点的潜在影响力降序选取节点，以避免由于随机顺序选取节点带来的社区结构划分不稳定现象；接着在计算待更新标签节点与其邻近节点的主题相似度和链接相似度基础上得出节点的综合相似度，并以其排序作为更新节点标签时对邻近节点遍历的顺序，保证了标签隶属度的稳定性，以提高生成社区的质量。

定义 1　社交网络 G。将社交网络抽象表示为一个有向图 $G(V,E)$，V 是 G 中用户节点的集合，E 是 G 中有向边的集合，其中 $V=\{v_1,v_2,\cdots v_n\}$，$|V|=n$，$E=\{\langle v,u\rangle \mid v,u\in V\}$，$\langle v,u\rangle$ 表示由节点 v 指向节点 u 的有向边。

初始化节点标签时对 G 中的每一个用户节点 v 赋予一个唯一的标签 c，表示其从属的社区。初始时指定所有标签的隶属度 $b(c,v)=1$。

2.1　标签传播过程

2.1.1 节点选择策略

COPRA 算法每次迭代都按随机顺序选择节点更新标签，生成的社区结构不稳定。MLPA-NCS 算法根据节点潜在影响力降序选择节点更新标签。

定义 2　节点潜在影响力 PI。节点潜在影响力 PI 表示节点在网络中的重要程度及对其他节点的影响程度。PageRank 中心性是有向网络特征向量中心性的变种，节点的中心性评估方法中特征向量中心性认为节点的重要性取决于邻接节点的度和邻接节点的重要性，因此可用 PageRank 中心性评估节点的

潜在影响力。G 中任意节点 v_i 的潜在影响力 PI 计算如式（3）所示。

$$PI(v_i)=(1-f)+f\sum_{v_j\in N(v_i)}\frac{PI(v_j)}{C_{DO}(v_j)} \tag{3}$$

其中：$N(v_i)$ 表示节点 v_i 的所有的父邻接节点组成的集合，$C_{DO}(v_j)$ 表示节点 v_j 的出度，f 表示阻尼系数，一般取 0.85。它是用来加速算法的收敛，而且可以避免由于孤立节点的存在而导致不能收敛的情况。

2.1.2 标签遍历顺序

确定了要更新节点的顺序后，就可以对所选节点更新其标签集。更新节点标签时以所选节点和邻近节点的综合相似度降序为标签遍历顺序，节点的综合相似度通过链接相关度[8]和主题相似度[9]计算得到。

邻近节点包括直接邻接节点和间接邻接节点。若节点 v,u 满足$(<v,u>\in E\vee<u,v>\in E)$，则称 v 和 u 互为直接邻接节点。若节点 v,u 满足$(<v,u>\notin E\wedge<u,v>\notin E\wedge((<v,w>\in E\wedge<u,w>\in E)\vee(<w,v>\in E\wedge<w,u>\in E)\vee(<v,w>\in E\wedge<w,u>\in E)\vee(<u,w>\in E\wedge<w,v>\in E)))$，则称 v 和 u 互为间接邻接节点。

定义 3 链接相关度 *link*。社交网络拓扑由用户节点和用户的双向关注构成，用户节点的链接相关度表示在网络拓扑中节点之间的链接紧密程度。用户节点 v 和 u 的链接相关度 *link* 定义如式（4）所示。

$$link_{vu}=\begin{cases}\frac{1}{2}adj_{vu}+\frac{1}{2}adj_{uv} & vu\text{直接邻接}\\ \alpha(co_{vu}+ci_{vu})+\beta S_{vu} & vu\text{间接邻接}\end{cases} \tag{4}$$

对$\forall x,y\in V$，若$<x,y>\in E$，则定义 x 到 y 的路径长度 adj_{xy} 为 1，否则为 0。当用户节点 v 和 u 互为间接邻接节点时，若$<v,w>\in E\wedge<u,w>\in E$，则表示 v 和 u 有共同指向关系，用 co_{vu} 表示；若$<w,v>\in E\wedge<w,u>\in E$，则表示 v 和 u 有共同被指向关系，用 ci_{vu} 表示；若$(<v,w>\in E\wedge<w,u>\in E)\vee(<u,w>\in E\wedge<w,v>\in E)$，则表示 v 和 u 有路径长度为 2 的链接关系，用 S_{vu} 表示。若$<v,w>\in E\wedge<u,w>\in E$，则 v 经 w 到 u 的路径长度 $Spl_{vu}=adj_{vw}+adj_{wu}=2$。$Spl_{uv}$ 也类似。间接邻接节点限定在路径长度为 2 的节点。式（4）中取 $\alpha=\beta=0.5$。co_{vu}、ci_{vu} 和 S_{vu} 如式（5）、（6）、（7）所示，O_v 表示节点 v 的出度，I_v 表示节点 v 的入度。

$$co_{vu}=\frac{|O_v\cap O_u|}{|O_v\cup O_u\cup I_v\cup I_u|} \tag{5}$$

$$ci_{vu}=\frac{|I_v\cap I_u|}{|O_v\cup O_u\cup I_v\cup I_u|} \tag{6}$$

$$S_{vu}=\frac{1}{2spl_{vu}}+\frac{1}{2spl_{uv}} \tag{7}$$

定义 4 主题相似度 *topic*。主题相似度用来衡量节点 v 与邻近节点 u 在主题上的相似程度。将用户的主题分布表示为向量空间的简单映射后，可通过主题概率分布计算得到两个用户的主题相似度。可用 KL（Kullback-Leibler divergence）距离的对称版本 JS（Jensen-Shannon）散度来衡量主题相异度，再根据主题相异度计算主题相似度。节点 v 和 u 之间的差异可用主题相异度公式 *dist(v,u)*[10]计算，如式（8）、（9）所示。若 V 为用户节点集，T 为主题集，则由 V 和 T 可构成"用户—主题"矩阵 UTM，UTM 反映了所有用户节点 v 的主题概率分布 VTP_v。式中 $D_{JS}(v,u)$ 是 VTP_v 和 VTP_u 之间的 JS 散度，$M=\frac{1}{2}(VTP_v+VTP_u)$ 是 VTP_v 和 VTP_u 的均值，$D_{KL}(P\|Q)=\sum_{t\in T}P(t)\log\frac{P(t)}{Q(t)}$ 是 Q 到 P 的 KL 散度。

$$D_{JS}(v,u)=\frac{1}{2}(D_{KL}(VTP_v\|M)+D_{KL}(VTP_u\|M)) \tag{8}$$

$$dist(v,u)=\sqrt{2\times D_{JS}(v,u)} \tag{9}$$

用户节点 v 和 u 的主题相似度定义如式（10）所示。

$$topic_{vu}=1-dist(v,u)=1-\sqrt{2\times D_{JS}(v,u)} \tag{10}$$

定义 5 综合相似度 sim_{vu}。融合节点 v 和 u 的主题相似度和链接相似度就得到用户的综合相似度，如式（11）所示。参数设置采用黄金分割比例，设 λ 数 0.618。

$$sim_{vu}=\lambda topic_{vu}+(1-\lambda)link_{vu} \tag{11}$$

2.1.3 更新节点标签

对待更新节点 v 的邻近节点根据综合相似度排序后，开始更新节点标签。节点接受邻近节点的标签影响程度与节点之间的综合相似度有关，综合相似度高的节点之间标签影响作用更明显，在此考虑将 sim_{vu} 值引入到节点标签的更新，v 的某个标签 c 在 t 轮迭代中的隶属度可用式（12）计算。

$$b_t(c,v)=\frac{\sum_{u\in N(v)}sim_{uv}\cdot b_{t-1}(c,u)}{|N(v)|} \tag{12}$$

其中：$N(v)$ 为节点 v 的邻近节点集合。

在更新节点标签时，如果某些节点携带无穷多隶属度较小的标签，使得该节点属于无穷多的社区，影响了所划分的社区的质量，所以需要设定淘汰机制。本文利用淘汰参数 δ（$0.27<\delta<0.62$）来限制节点所拥有的标签个数，根据经验本文取其值为 0.6。在每次标签更新完成后，对隶属度做归一化处理，使得每一个节点所拥有的标签隶属度总和为 1，将标签集合中的元素根据隶属度进行降序排序，从隶属度最大的值开始累加，直至和不小于 δ，选取这前几个标签并重新做归一化处理，使得 $\sum_{c\in label(v)}b(c,v)=1$。

2.2 MLPA-NCS 算法描述

输入：网络 $G(V,E)$，用户节点集 V，主题 T，淘汰参数 δ，每个节点的标签集 *label*，标签集初始为空。

输出：重叠社区集合 C。

1.初始化节点标签，为 G 中每个节点 v 赋予唯一的标签 c；

2.由用户节点集 V 和主题集 T 构造“用户—主题”矩阵 UTM，并计算每个节点 v 的主题概率分布 VTP_v。

3.据式（3）计算所有节点 v 的潜在影响力 $PI(v)$，并由高到低排序；

4.令 t=1；

5.按潜在影响力 PI 降序更新所有节点 v 的标签，重复执行（5.1）~（5.6）步:

5.1 据式（4）计算 v 与其所有邻近节点 u 的链接相关度 $link_{vu}$。

5.2 据式（10）计算 v 与其所有邻近节点 u 的主题相关度 $topic_{vu}$。

5.3 据式（11）计算 v 与其所有邻近节点 u 的综合相似度 sim_{vu}。

5.4 更新 v 的标签，即记录 v 接受到的所有邻近节点 u 的标签 c。在根据 sim_{vu} 对邻近节点进行降序排序的基础上，利用隶属度将每一个邻近节点 u 的所有标签排序，以此作为计算 v 接受到所有标签的隶属度的顺序，按式（12）计算每一个标签 c 的隶属度 $b_t(c,v)$。

5.5 对 v 的所有标签的隶属度 b 进行初步归一化处理，使得 $\sum_{c\in label(v)} b_t(c,v)=1$；

5.6 根据淘汰参数 δ 对 v 的初步归一化结果进行过滤，过滤之后进行二次归一化处理；

6.如果 G 中所有节点 v 的标签集 $label(v)$不再改变，标签传播过程停止，转到第 7 步；否则，令 $t=t+1$并转到第 5 步；

7.根据所有节点最终的标签集确定其所属社区，得到社区集合 $C=\{C_1,C_2,\ldots,C_q\}$。

2.3 算法复杂度分析

假定网络有 n 个节点和 m 条边，m/n 表示节点的平均邻居数。

a）初始化节点标签需要时间复杂度 $O(n)$；

b）计算所有节点的主题概率分布需要时间复杂度 $O(n)$；

c）根据 PageRank 计算节点的潜在影响力需要时间复杂度 $O(n\log n)$；

d）计算节点与邻近节点的综合相似度需要时间复杂度 $O(m)$；

e）与 COPRA 算法类似，每个节点接受其每个邻居节点标签的时间复杂度同为 $O(\log(m/n))$，这一阶段总的时间复杂度为 $O(m\log(m/n))$。

根据以上分析，忽略掉较小的时间复杂度，MLPA-NCS 算法总的算法复杂度为 $O(m\log(m/n))$，比 COPRA 算法的时间复杂度 $O(vm\log(vm/n))$略低。

3 实验

为了考察本文提出的 MLPA-NCS 算法的可行性，并且验证其相比现有的基于多标签传播的重叠社区划分算法划分的社区具有更高的质量和准确性，本文采用了真实数据集和人工网络图进行实验，对 LPA，COPRA，SLPA，LPPB 和 MLPA-NCS 算法进行对比。所有算法和数据的运行环境为 Core i5-2450M，12GB，Microsoft Windows10，在 anaconda2 平台上进行实验。

3.1 实验数据

实验采用的真实数据集是 Karate[11]，该数据集是目前社区划分研究中使用的小型复杂网络数据集。该数据集描述美国一个空手道俱乐部的成员关系，网络包含 34 个节点和 78 条边，实际可划分为两个社区结构，如表 1 所示。

表 1　Karate 网络实际分组

社区	成员编号
1	1 2 3 4 5 6 7 8 11 12 13 14 17 18 20 22
2	9 10 15 16 19 21 23 24 25 26 27 28 29 30 31 32 33 34

由于 Karate 数据量较少，社区划分结果可直接与网络实际分组进行比较，从而验证算法的可行性。对于大型网络，则采用 LFR 网络生成程序仿真生成较大规模的人工网络图，本文采用 LFR-10000 人工网络对 LPA、COPRA、SLPA、LPPB、MLPA-NCS 五种算法进行对比。LFR benchmark 基准程序由 Lancichinetti 等人[12]提出，根据参数设置生成所需求的网络，本实验网络参数如表 2 所示，其中，N 为节点数目，k 为节点平均度数，maxk 为节点最大度数，minc 为社区最小规模，maxc 为社区最大规模，mu（mixing parameter）为节点与社区外部连接的边数与该节点度数的比值，该比值越小，说明节点可连接的社区越少，网络的社区结构越明显，mu 取 0.1～0. 6，每次增加 0.05，生成 11 个 LFR-10000 网络。

表 2　LFR-10000 人工网络参数设置

参数	LFR-10000
N	10000
k	10
maxk	300
minc	30
maxc	100
mu	0.1~0.6

3.2 实验评价标准

实验采用标准化互信息 NMI[13]度量社区划分算法生成的社区结构与标准社区结构之间的相关性，以此评估算法的准确性，如式（13）所示。采用重叠模块度 Qov[14][15]评价重叠社区的网络结构，以此度量社区划分的质量，如式（14）~（16）所示。

$$NMI(X|Y)=1-[\frac{1}{2}(H(X|Y)_{norm}+H(Y|X)_{norm})] \quad (13)$$

其中：X 和 Y 分别代表实验划分的社区结构和标准的社区结构。NMI 值越大说明划分结果与标准网络结构越相似，算法划分社区的准确性越高。

$$Qov=\frac{1}{m}\sum_{c\in C}\sum_{i,j\in V}[r_{ijc}A_{ij}-\omega_{ijc}\frac{k_i^{out}k_j^{in}}{m}] \tag{14}$$

其中：A为邻接矩阵，K为节点的度，m为边的个数，r_{ijc}表示节点 i 和 j 同属于社区 c 的概率，$r_{ijc}=\ell(p_{i,c},p_{j,c})$，$p_{i,c}$表示 i 属于社区 c 的概率，ω_{ijc}表示节点 i 或者节点 j 在社区 c 中的概率。

$$\ell(p_{i,c},p_{j,c})=\frac{1}{(1+e^{-f(p_{i,c})})(1+e^{-f(p_{j,c})})} \tag{15}$$

$$\omega_{ijc}=\frac{\sum_{j\in V}\ell(p_{i,c},p_{j,c})}{|V|}\times\frac{\sum_{i\in V}\ell(p_{i,c},p_{j,c})}{|V|} \tag{16}$$

根据文献[17]建议，本文 f 定义为 $f(x)=60x-30$。Qov 取值范围在 0 到 1 之间，值越大，重叠社区结构越好。

3.3　实验结果与分析

实验中对 Karate 数据集用 MLPA-NCS 进行社区划分后的结果如图 1 所示，可以看出共分为两个社区，与表 1 中实际分组相比，除节点 3 和 31 为两个社区的重叠节点外，其他节点都分配到各自所属的社区且均划分正确，取得了较好的实验效果。

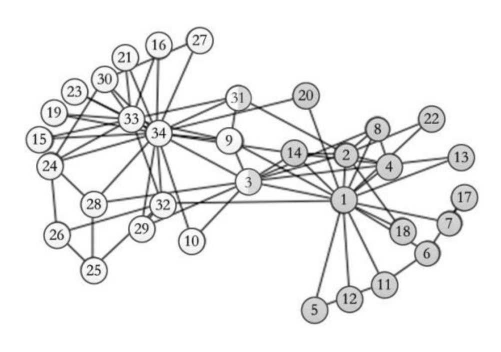


图 1　MLPA-NCS 对 Karate 网络的划分结果

实验选用 LPA、COPRA、SLPA、LPPB、MLPA-NCS 五种算法作为对比，为避免算法的随机性对实验造成影响，对 LPA、COPRA、SLPA、LPPB 四个算法都进行 20 次实验取结果平均值，而本文提出的 MLPA-NCS 由于算法的稳定性只需进行一次实验。采用 LFR 网络生成程序仿真生成不同规模的人工网络图，求取各算法 NMI 与 Qov。图 2 和 3 是在 LFR-10000 的人工网络中，五种算法划分的社区结构随着 mixing parameter 的改变所求得 NMI 与 Qov 的变化情况。

从图 2 可知，在 mixing parameter 值较小的时候，网络的社区结构较为明显，社区之间的边界较为清晰，此时 LPPB 与 MLPA-NCS 算法相比其他三个算法的 NMI 值较高，但是随着 mixing parameter 值的增大，MLPA-NCS 算法的 NMI 值超过 LPPB 算法，由此可以判断大规模网络中 MLPA-NCS 算法准确性更高。从图 3 可知，无论 mixing parameter 值是大或小，均可以看出 LPPB 与 MLPA-NCS 算法划分的社区模块度 Qov 比其他三个算法更有优势，原因是这两个算法对标签传播算法的随机策略的改进减少了运行结果的差异，在一定程度上发挥出潜在影响力更高的节点的作用，实验结果表明 MLPA-NCS 算法无论是在社区结构明显还是模糊的网络中，都能划分出质量较高的社区。

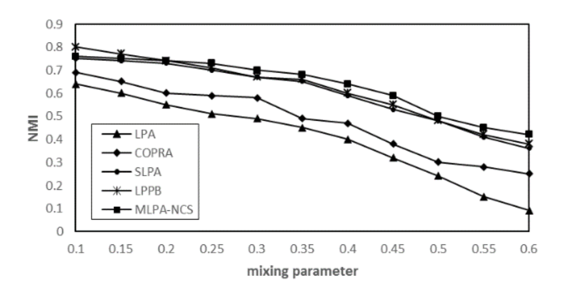


图 2　算法在 LFR-10000 上 NMI 的比较

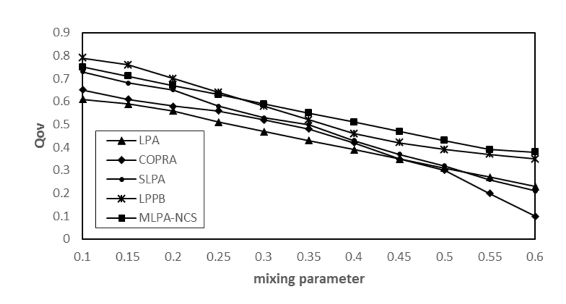


图 3　算法在 LFR-10000 上 Qov 的比较

总体而言，MLPA-NCS 算法相比现有的多标签传播社区划分算法在一定程度上提高了社区划分的稳定性和生成社区的质量。

4　结束语

针对 COPRA 算法在选择节点和节点标签更新时因采用随机顺序策略导致传播过程不确定、社区划分结果不稳定且生成社区质量不够高等问题，本文以 COPRA 算法框架为基础对其改进，提出基于节点综合相似度的多标签传播社区划分算法 MLPA-NCS 算法，通过计算节点潜在影响力并排序作为选择更新节点的顺序，通过计算由节点主题相似度和链接相关度构成的综合相似度并排序作为节点标签更新的顺序，提高了生成社区的质量，保证了社区划分结果的稳定。实验表明 MLPA-NCS 算法在 Karate 数据集上的社区划分结果正确且有较好的实验效果，同时在 LFR-10000 上的实验表明当 mu 值比较大时，MLPA-NCS 算法相比于 LPA、COPRA、SLPA、LPPB 四个算法的 NMI 值较高，说明 MLPA-NCS 算法准确性更高；相比于 LPA、COPRA、SLPA、LPPB 四个算法划分的社区结构具有更高的模块度 Qov，说明社区质量相对更高且社区划分结果稳定。

参考文献：

[1] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical Review E: Statistical Nonlinear & Soft Matter Physics, 2007, 76 (2): 036106.

[2] Barber M J, Clark J W. Detecting network communities by propagating labels under constraints [J]. Physical Review E: Statistical Nonlinear & Soft

Matter Physics, 2009, 80 (2 Pt 2): 026129.

[3] Gregory S. Finding overlapping communities in networks by label propagation [J]. New Journal of Physics, 2009, 12 (10): 2011-2024.

[4] Xie J, Kelley S, Szymanski B K. Overlapping community detection in networks: The state-of-the-art and comparative study [J]. ACM Computing Surveys, 2013, 45 (4): 43.

[5] 刘世超，朱福喜，甘琳．基于标签传播概率的重叠社区发现算法 [J]. 计算机学报, 2016, 39 (4): 717-729. (Liu Shichao, ZhuFuxi, Gan Lin. A label-propagation-probability-based algorithm for overlapping community detection [J]. Chinese Journal of Computers, 2016, 39 (4): 717-729.)

[6] 张昌理，王一蕾，吴英杰，等．基于信息熵和局部相关性的多标签传播重叠社区发现算法 [J]．小型微型计算机系统, 2016, 37 (8): 1645-1650. (Zhang Changli, Wang Yolei, Wu Yingjie, *et al.* Multi-label propagation algorithm for overlapping community discovery based on information entropy and local correlation [J]. Journal of Chinese Computer Systems, 2016, 37 (8): 1645-1650.)

[7] Gui Q, Deng R, Xue P, *et al.* A community discovery algorithm based on boundary nodes and label propagation [J]. Pattern Recognition Letters, 2017.

[8] 闫光辉，舒昕，马志程，等．基于主题和链接分析的微博社区发现算法 [J]．计算机应用研究, 2013, 30 (7): 1953-1957. (Yan Guanghui, Shu Xin, Ma Zhicheng, *et al.* Community discovery for microblog based on topic and link analysis [J]. Application Research of Computers, 2013, 30 (7): 1953-1957.)

[9] Endres D M, Schindelin J E. A new metric for probability distributions [J]. IEEE Trans on Information Theory, 2003, 49 (7): 1858-1860.

[10] Weng J, Lim E P, Jiang J, *et al.* TwitterRank: finding topic-sensitive influential twitterers. [C]// Proc of the 3rd ACM International Conference on Web Search and Data Mining. 2010: 261-270.

[11] Zachary W W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977, 33 (4): 452-473.

[12] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms [J]. Physical Review E: Statistical Nonlinear & Soft Matter Physics, 2008, 78 (4 Pt 2): 046110.

[13] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure of complex networks [J]. New Journal of Physics, 2008, 11 (3): 19-44.

[14] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E: Statistical Nonlinear & Soft Matter Physics, 2004, 69 (2 Pt 2): 026113.

[15] Nicosia V, Mangioni G, Carchiolo V, *et al.* Extending the definition of modularity to directed graphs with overlapping communities [J]. Journal of Statistical Mechanics Theory & Experiment, 2008, 2009 (3): 3166 – 3168.